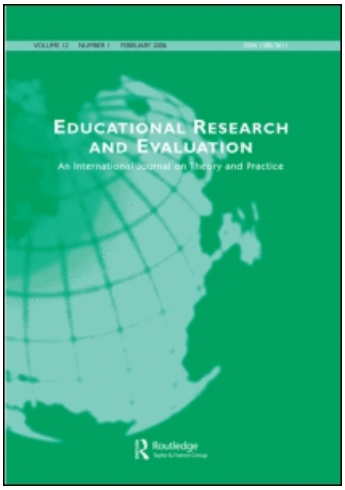
Literacy programs for initial reading instruction: Do they make a difference in learning outcomes. [IF ..]

Blok, H.; Otter, M.E.; Overmaat, M.; de Glopper, K.; Hoeksma, J.B.

published in
Educational Research and Evaluation
2003

DOI (link to publisher)
10.1076/edre.9.4.357.17810

document version
Publisher's PDF, also known as Version of record



Educational Research and Evaluation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t714592776

Literacy Programs for Initial Reading Instruction: Do They Make a Difference in Learning Outcomes?

Henk Blok; Martha E. Otter; Marianne Overmaat; Kees de Glopper; Jan B. Hoeksma

Online publication date: 09 August 2010

To cite this Article Blok, Henk , Otter, Martha E. , Overmaat, Marianne , de Glopper, Kees and Hoeksma, Jan B.(2003) 'Literacy Programs for Initial Reading Instruction: Do They Make a Difference in Learning Outcomes?', Educational Research and Evaluation, 9: 4, 357 — 371
To link to this Article: DOI: 10.1076/edre.9.4.357.17810
URL: http://dx.doi.org/10.1076/edre.9.4.357.17810

Educational Research and Evaluation
2003, Vol. 9, No. 4, pp. 357–371

1380-3611/03/0904–357$16.00

Literacy Programs for Initial Reading Instruction: Do They Make a Difference in Learning Outcomes?

Henk Blok[1], Martha E. Otter[1], Marianne Overmaat[1], Kees de Glopper[2], and Jan B. Hoeksma[3]
[1]University of Amsterdam, The Netherlands, [2]University of Groningen, The Netherlands, and [3]Vrije Universiteit, Amsterdam, The Netherlands

ABSTRACT

Most teachers in developed societies use ready-made programs to teach initial reading. The present study compares the effects of 3 such programs, covering 70% of the market. Programs differed with respect to their student versus whole-class orientation and the availability of teaching materials. The program outcomes investigated were Word Reading, Reading Comprehension, and Spelling. In addition, bias against immigrants or gender was tested. The sample consisted of 425 students (mean age 78 months, SD = 4.5 months) from 46 different schools. Data were analyzed by means of the hierarchical linear model. Results showed that the programs did not differentially affect mean outcomes. Nor did they affect the variability between schools and between students or bias against specific groups. It is concluded that the choice between these 3 particular reading programs has only marginal consequences for learning outcomes. It is suggested that future research should be directed at effects of literacy programs on interindividual variability of pupils.

INTRODUCTION

Many teachers rely on basal readers to teach initial reading. Literacy programs, or reading schemes as they are called in the United Kingdom, relieve teachers in many ways. They basically serve as curricula written out in considerable

Address correspondence to: Henk Blok, University of Amsterdam, Faculty of Social and Behavioural Sciences, Department of Education, SCO-Kohnstamm Instituut, Wibautstraat 4, 1091 GM Amsterdam, The Netherlands. E-mail: henkb@educ.uva.nl

Manuscript submitted: October 18, 2001.
Accepted for publication: May 16, 2002.

detail. They do not only offer suggestions for day-to-day activities, but also contain practical materials for teaching and learning, such as word lists, work sheets, suitable texts, courseware or instructive games. Thanks to the literacy program, teachers need less time to prepare lessons. Furthermore, many programs contain detailed planning schedules that facilitate cooperation between teachers. The latter may be especially important when students are instructed by more than one teacher, for instance when two part-time teachers share the same class.

Figures about the percentages of teachers using literacy programs are not readily available. It does not seem a bold assumption, however, that the vast majority of teachers in developed countries use published literacy programs of some sort. In The Netherlands, there has been a long-standing tradition to teach reading and writing by means of literacy programs. A nationwide survey in 1996/1997 showed that about 90% of the grade-1 teachers relied heavily on published literacy programs (Overmaat & Ledoux, 1998). Most of these programs offer essentially all the materials needed by teachers and students, including comprehensive teacher's manuals, graded series of books and work sheets for students, and diverse materials such as boxes with separate letters, wall charts, et cetera.

Similar figures are available for the United States. In the 1980s, the use of basal readers accounted for approximately 90 to 95% of all reading instruction (Goodman, Shannon, Freeman, & Murphy, 1988). Although the exclusive use of basal readers has undergone a dramatic shift since then (Martinez & McGee, 2000), still 83% of the U.S. teachers reported that they either use basals supplemented by children's books or children's books supplemented by basals (Baumann, Hoffman, Moon, & Duffy-Hester, 1998).

The goal of the present study was to compare the outcomes of three different literacy programs commonly used in The Netherlands. It was examined whether students that learn to read and write with one program reach better results – in terms of reading and writing ability – than students working with another program. The answer to this question is not only relevant for schools and teachers who consider the replacement of their literacy program. Also educational authorities, many of whom try to strengthen their control of educational productivity, want to know whether differences exist between programs and, if so, which programs are the more effective ones.

REVIEW OF RELATED RESEARCH

Given the overwhelming use of literacy programs, it is surprising that virtually no studies have been published comparing the differential effectiveness of such programs. For sure, not every program leads to fully satisfying results. Estimates of the number of students experiencing reading failure remain high throughout the years. Stedman and Kaestle (1987) reported estimates of reading failure to be as high as 20 to 25%. A summary of data from the 1994 National Assessment of Educational Progress revealed that 44% of U.S. fourth graders scored ''below basic'' on the reading assessment (Williams, Reese, Campbell, Mazzeo, & Phillips, 1995). Dutch figures seem less dramatic, but still 10 to 15% of the students in grade 1 experience failure in learning to read and are at risk for retention or referral to special education.

Possibly because of the famous *First grade studies* by Bond and Dykstra (1967) many researchers gave up to compare methods (Searfoss, 1997). Bond and Dykstra recommended that ''Future research might well center on teacher and learning characteristics rather than method and materials'' (1967/1997, pp. 123/415–416). Searfoss therefore suspects that researchers ''have made it passé and anti-intellectual, at times, for teachers to ask the research community to help them find out what works'' (1997, p. 437). Even today teachers are told that the search for a perfect method is illusive (Duffy & Hoffman, 1999). The latter authors reject the perfect method concept, favoring ''the development of teachers who know a variety of methods and approaches, and who orchestrate those thoughtfully and adaptively according to their students' needs'' (p. 13). As noted in *Preventing Reading Difficulties in Young Children* (Snow, Burns, & Griffin, 1998), there is ''... currently no requirement and little incentive for publishers or adopting schools to evaluate reading-related materials and in-service programs in terms of efficiency'' (p. 333). Snow et al. recommended that ''... materials purveyors that currently do not provide adequate evidence to support data-based decision making about their products should be required to do so'' (p. 334).

Studies comparing programs are rare indeed. In fact, we found only a few studies comparing reading or literacy programs. Stein, Johnson, and Gutlohn (1999) selected seven first-grade basal reading programs adopted by California in 1996. From available research they derived two instructional features characteristic of effective reading programs. The first feature was whether *phonics* are instructed explicitly or implicitly. In explicit phonics instruction, the phonemes associated with graphemes are identified in isolation and next

blended together into words. In implicit phonics instruction, students are asked to identify the phonemes in the context of whole words, rather than in isolation. The second instructional feature involved the *amount of decodable text*, that is text that conforms to the grapheme-phoneme correspondences that were previously introduced. Based on an analysis of curricular content, the authors found sizeable differences between the reading programs. Only one program used an explicit phonics approach. With respect to the second instructional feature, the authors concluded that in six out of seven programs the amount of decodable text was small (less than 15% wholly decodable words). Considering what these authors believe to be effective program characteristics (explicit phonics, combined with a high percentage of wholly decodable words), most of the programs are expected to result in sub-optimal learning outcomes.

If reading instruction affects students achievement, an assumption that does not seem very audacious, one might expect differential learning outcomes between reading programs differing with respect to essential instructional features. Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998) demonstrated differential effects indeed. They compared the learning outcomes of three different kinds of classroom reading programs. The three programs involved direct instruction in letter-sound correspondences (direct code), less direct code instruction embedded in connected text (embedded code), and implicit code instruction (implicit code). Children receiving direct code instruction reached better scores for word recognition and reading comprehension, although the latter effect was less robust. No differences were shown for spelling ability. The differences found are in line with current views on the importance of explicit instruction in the alphabetic principle for at-risk children. However, Foorman et al. (1997) examined learning outcomes of three different ''off-the-shelf programs'' in a sample of reading disabled students, but no differences were found.

In The Netherlands, several studies have been done comparing literacy programs on the basis of learning outcomes. Because these studies were published only in Dutch, their impact on the international research community has been limited. The Dutch studies generally lead to the conclusion that there are weak relationships at best between different literacy programs and learning outcomes. The study by Kooreman (1974) offers a notable exception. It demonstrated a mean difference between two literacy programs of about 0.8 standard deviation on a measure of word identification accuracy. According to Cohen's criteria, a difference of this size should be considered large (Cohen, 1988).

In The Netherlands, as in many other countries, it is not only the general effectiveness question that matters. Educational officials are anxious about the possibility that some literacy programs are biased against specific subpopulations. The present study focuses on two different student characteristics that have known negative effects on reading acquisition. The first characteristic is whether the student or his/her parents are immigrants. Many immigrant students lag behind in reading and writing abilities. In general, immigrant students receive less home support than students born from Dutch families. In addition, and this may be more important, immigrant students are less proficient in the instruction language, which is exclusively Dutch. The second possible bias pertains to gender. Dutch girls consistently outperform boys in a variety of reading tasks, including word reading and reading comprehension. Differences range from 0.2 to 0.5 standard deviation, as was established in a nationwide sample of 8-year-old students (Sijtstra, 1992). In the same sample, the disadvantage of immigrant students amounted to 0.3 standard deviation, for word reading and reading comprehension. Disadvantages in reading for boys and immigrant students are demonstrated in nearly every participating country in the ''Reading Literacy Study'' of the International Association for the Evaluation of Educational Achievement (Elley, 1992).

In sum, the present article not only pertains to general differences in learning outcomes, but also to differences in bias against immigrant students and boys. Leading question is whether these differences are related to the literacy program in use.

METHOD

Literacy Programs

We compared three different published literacy programs: ''Every Child a Reader'' (Dutch: Alle kinderen leren lezen), ''Roads to Reading'' (Dutch: Leeslijn), and ''Risk-Free Reading'' (Dutch: Veilig leren lezen). Together the three programs cover nearly 70% of the literacy programs used in The Netherlands (''Every Child a Reader'' covers less than 1%, ''Roads to Reading'' 8% and ''Risk-Free Reading'' 61%). ''Every Child a Reader'', although rarely used, was selected because the program represents the belief that teachers need only basic suggestions and materials. Comprehensive programs carry the risk of deskilling teachers (cf. Baumann & Heubach,

1996). In addition, this program is remarkably cheap. It sells for less than 20% of the price of the other two literacy programs. This would make ''Every Child a Reader'' very attractive, if its effectiveness could be demonstrated.

A common characteristic to these three programs is the adherence to an explicit phonic approach. Students learn to manipulate phonemes in a variety of tasks. In The Netherlands, whole-language approaches are not seen as a viable way of reading instruction. Indeed, research shows that the instruction of phonemic awareness has a beneficial effect on reading acquisition (Ehri et al., 2001).

Every Child a Reader (ECR) is a very simple literacy program. Its intended simplicity emerges from the appearance of student materials (only black print, almost no pictures), the succinctness of teacher materials (a document comprising less than 30 pages), and the lack of graded reading series and of remedial exercises. Instructional suggestions for day-to-day activities are kept to a minimum. ECR favors whole-class instruction, assuming that students entrance levels are more or less the same. A special feature of the program is that it contains an optional kindergarten course, directed at phonological skills (rhyming, blending of phonemes, and segmenting of words).

Roads to Reading (RtR) covers the reading curriculum from emergent literacy to fluent text reading, as typically taught from kindergarten until grade 4 (reading comprehension is not included). RtR favors *individual* instruction, as opposed to whole-class instruction. Teachers are advised to organize their class into subgroups of students needing more or less explicit instruction and exercises. RtR offers two separate courses, one following principles of discovery learning, the other following principles of direct instruction. In the latter case the program is highly structured. The materials of RtR include work sheets, graded reading series and reading games. The teacher manuals contain more than 500 pages.

Risk-free Reading (RfR), incorporating a course for emergent literacy, covers the reading curriculum from kindergarten until grade 3. It comprises reading, including reading comprehension and spelling. Both domains are integrated from the beginning. Although RfR favors whole-class instruction, the program offers many opportunities for students to work at their own level. Attractive computer games, work sheets and books, combined with audio tapes, enhance students' possibilities for autonomous learning. The teacher materials are very comprehensive, as in RtR.

Participants

Address files were obtained from the three publishers who keep records of the schools they sell their literacy programs to. Twenty schools were randomly sampled from each file. To spare teachers excessive test chores, it was decided to sample randomly 10 students within each school. We estimated that samples of 20 schools and 10 students per school would be large enough to detect differences of 0.5 standard deviation. Complete data were available for 425 students, stemming from 46 schools. Table 1 describes the composition of sample. It shows that the ECR group is smaller than intended. This is due to the fact that the population of schools using ECR is small. Within the ECR population we could not find enough schools willingly to participate. There was some further attrition of students in each subsample, because tests were taken at the start and at the end of the school year. Illness or removals of students accounted for a small amount of missing data (between 7 and 8%). The mean age of students at the start of the study was 78 months (SD is 4.5 months).

Variables

The literacy programs for initial reading instruction were compared with respect to their effects on Word Reading, Reading Comprehension, and Spelling Ability. Word Reading was taken individually, the remaining tests were group administered. These tests were administered by teachers, according to strict procedures, by the end of the school year of grade 1. To test for initial differences between schools, a Vocabulary Test was given at the beginning of the school year, that is between 6 and 8 weeks after schools started. All tests used are standardized tests developed by Cito (the Dutch National Institute for Test Development) and widely in use.

Table 1. Composition of the Sample.

Literacy program	Number of schools	Number of students	Percentage of boys	Percentage of immigrant students	Mean (s.d.) Vocabulary Test
ECR	10	89	48	9	35.7 (6.5)
RtR	17	159	44	11	36.0 (6.9)
RfR	19	177	48	10	35.5 (7.6)
Total	46	425	47	10	35.7 (7.1)

Vocabulary was measured by a test for receptive vocabulary. Students match words with pictures by choosing between four different pictures. The test contains 50 items. The score is the number correct. This test was used as a measure of general verbal ability to test for pre-existing differences among the three subsamples.

Word Reading was measured with three word lists, with increasing difficulty. The first list contained words build from three phonemes, consonant-vocal-consonant (CVC). The second list contained words build as CCVC, CVCC, or CCVCC. The third list contained multisyllable words, still more difficult to read. Students were allowed 1 min for each list to read as many words as possible. The raw score is the total number of words read correctly over the three lists. The resulting score distribution appeared to be positively skewed, due to students reaching very high scores. Therefore the scores were transformed by taking the square root.

Reading Comprehension consisted of several short passages followed by multiple choice questions. The test totals 51 items. The raw score (the number of items answered correctly) was transformed according to a two parameter logistic model (Verhelst & Glas, 1993) that allows for the use of different test forms or booklets.

Spelling Ability was measured by a dictation task containing 37 words. The words are presented in a sentence context. The raw score (number of words written correctly) was again transformed following a two parameter logistic model (Verhelst & Glas, 1993) that allows for different test forms or booklets.

Data Analysis

Because students were nested in schools, data were analyzed using the hierarchical linear model (Bryk & Raudenbush, 1992; Goldstein, 1995). Students make up the first level, and schools the second level. Separate analyses were performed for each of the three dependent variables, Word Reading, Reading Comprehension, and Spelling Ability. Effects of the literacy programs, gender and immigrant status, and their possible interactions were entered by means of dummy variables. The so-called level-2 variance corresponds to random differences between schools. The level-1 variance is interpreted as random variability within schools. Significance was tested using the likelihood ratio test (Goldstein, 1995). The corresponding test statistic designated by G^2 follows a chi-square distribution. The number of degrees of freedom (*df*) equals the number of dummy variables entered.

RESULTS

The main question of the present study pertains to differences between the three literacy programs on Word Reading, Reading Comprehension, and Spelling Ability. Initial differences between schools with respect to the percentages of boys and girls, percentage of immigrants, and vocabulary ability (as a proxy of verbal ability) are possible confounders of the program effects. Table 1 shows the percentages of boys, percentages of immigrants, and mean vocabulary size for each literacy program. All differences were small and proved to be nonsignificant ($p > .10$).

It was hypothesized that literacy programs may not only affect the mean level of abilities (i.e., Word Reading, Reading Comprehension, and Spelling Ability), but also the random differences between schools (the level-2 variance) and random differences within schools (the level-1 variance). The latter would be the case if the better students within schools would profit from one of the programs, whereas the same program would be less favorable for disadvantaged students. Table 2 displays the estimated means, between-school (level-2) and within-school (level-1) variances of the three programs for each ability. A closer look at the within- and between-school variances in Table 2 reveals a systematic pattern. For RtR the within-school variances (i.e., the differences between pupils) are relatively large, whereas the corresponding between-school variances (differences between schools) are relatively small.

The differences between variances across programs were tested next, by comparing the likelihood of the model with a common within and a common between variance to the models with program-specific variances. In spite of the systematic pattern, the final row section of Table 2 shows that the differences were not significant. In other words, the programs do not differentially affect the variability of Word Reading, Reading Comprehension, or Spelling Ability. Given these results, all subsequent analyses assumed a common level-2 and a common level-1 variance across the three programs.

The next analyses were aimed at finding fixed (i.e., nonrandom) differences between the three literacy programs. Table 2 already contained the estimated means (and standard errors) of the three literacy programs. The differences were small. Expressed in mean effect sizes they amount to 0.20 standard deviation for both Word Reading and Reading Comprehension and 0.25 standard deviation Spelling Ability. Note that Cohen (1988) would classify these as small. Comparing the means using the likelihood ratio test gave nonsignificant results for each of the abilities. Table 3 contains the

Table 2. Models With Separate Level-1 and Level-2 Variances; Means (Standard Errors), Between-Schools Variances (Standard Errors) and Within-Schools Variances (Standard Errors).

Dependent variable/literacy program	Mean	Stand. error	Between-schools		Within-schools	
			Variance	Stand. error	Variance	Stand. error
Word Reading						
ECR	8.52	0.41	1.03	0.77	5.82	0.93
RtR	9.32	0.24	0.30	0.34	6.11	0.72
RfR	9.20	0.35	1.63	0.74	5.79	0.65
Reading Comprehension						
ECR	191.64	2.60	26.07	30.63	363.28	57.77
RtR	195.47	1.88	18.19	21.10	389.37	46.19
RfR	198.15	2.79	109.80	48.23	345.31	38.84
Spelling Ability						
ECR	109.09	1.61	22.58	11.65	29.82	4.75
RtR	110.45	0.70	4.30	2.91	37.84	4.49
RfR	111.78	0.98	14.30	5.97	36.31	4.08
Test for equal variances (df = 4)						
Word Reading			$G^2 = 3.15$ (n.s.)			
Reading Comprehension			$G^2 = 4.65$ (n.s.)			
Spelling Ability			$G^2 = 5.95$ (n.s.)			

Table 3. Results of Hypothesis Tests Concerning Equal Means and Bias Against Gender or Immigrant Children.

Hypothesis/dependent variable	Df	G^2	P-level
Equal means			
Word Reading	2	2.53	n.s.
Reading Comprehension	2	2.99	n.s.
Spelling Ability	2	2.84	n.s.
No gender bias			
Word Reading	2	0.59	n.s.
Reading Comprehension	2	0.45	n.s.
Spelling Ability	2	1.77	n.s.
No immigrant status bias			
Word Reading	2	1.45	n.s.
Reading Comprehension	2	0.11	n.s.
Spelling Ability	2	3.50	n.s.

corresponding test statistics. The different programs do not have different effects on the mean level of Reading Ability, Reading Comprehension, and Spelling Ability.

Bias of any of the programs against immigrant children was tested by comparing two models, one model containing different means for each program and three additional interaction terms (defined by multiplying the three program dummies with immigrant status), and the simpler model containing only three program-specific means and one general effect for being an immigrant. The more elaborate model did not result in a better fit to the data, for any of the dependent variables (Table 3). In other words, the programs did not have differential effects with respect to immigrant status. For two of the three abilities the betas or regression weights indicate that immigrants perform less well than nonimmigrants. For Reading Comprehension beta amounted to −14.28 (standard error 3.58), for Spelling Ability beta amounted to −4.44 (standard error 1.24).

The final analyses pertained to gender differences. Testing proceeded in the same way. The more elaborate models, containing three interaction effects (''program dummies by gender'') did not fit better than the models with only one general gender effect, indicating that the programs demonstrate no differential gender bias (Table 3). The mean effects for girls were positive and significant for two of the three abilities. For Word Reading beta amounted to 0.59 (standard error 0.25), for Reading Comprehension beta amounted to 4.33 (standard error 1.90).

CONCLUSIONS AND DISCUSSION

The analyses did not reveal any important statistically significant differences between the learning outcomes of the three literacy programs. In other words, the programs examined do not differentially influence the performance of the pupils with respect to Word Reading, Reading Comprehension, and Spelling Ability. Nor do the programs affect student variability within and between schools. No differential bias was observed against immigrant children or boys either.

These results should be seen as very reassuring for both teachers and educational officials. The choice between the involved literacy programs apparently does not have important consequences for the learning outcomes of pupils, even when literacy programs are as different as they were in the present

study. It is also of practical importance that the choice of a specific program does not enhance differences between immigrant and native pupils and between boys and girls.

Learning outcomes are only one part of the story, however. Not every teacher is equally satisfied with his or her literacy program. In an additional part of the present study, not extensively reported here, teachers were asked to rate the literacy programs they used, on six general criteria derived from 37 specific questions (for details, consult Blok, Otter, & De Glopper, 2000). Users of RfR were content on all criteria, but users of ECR and RtR were dissatisfied in one or more respects. ECR users were particularly negative about the attractiveness of student materials, whereas RtR users were negative about the organization of the teacher manuals, more specifically about the day-to-day suggestions. As the present study shows, these criticisms apparently have no consequences for the learning outcomes.

The present article pays limited attention to program compliancy. Our study deliberately evaluated the effects of the teachers and students working with the programs in natural use, rather than the effects of the programs in strictly controlled and therefore artificial contexts. Still, we have some additional data available regarding treatment fidelity. At the end of the school year, teachers were questioned through a 20-item questionnaire, about different aspects of program compliancy. Most teachers indicated that they followed their program quite strictly. Only ECR teachers deviated relatively often by supplementing their program with elements from other literacy programs or with self-designed materials, for instance work sheets. Knowing the specific character of the ECR program, offering only the very basic materials, this finding should be hardly surprising. In fact, in the manual ECR teachers are encouraged to supplement the program with whatever materials or exercises they need. For this reason one might argue that program compliancy was generally high, even in the ECR condition.

An important question is whether the absence of statistical significant differences between the outcomes could be due to a statistical artifact, more specifically small sample size. The sample was planned on the assumption that it could reveal medium-sized differences, that is differences of approximately 0.5 standard deviation. The estimated models allow calculating the a posteriori power to detect medium-sized differences. Using Snijders and Bosker (1999, p. 142) and the estimated standard errors of the means in Table 2, we calculated the a posteriori power to reveal significant differences between two programs. For both Word Reading and Reading Comprehension, the power appeared to be

well above $p > .90$, whereas for Spelling Ability it was less than $p < .50$. In other words, in the present sample the probability to detect a medium-sized difference was over 90% for Word Reading and Reading Comprehension. It is unlikely that the absence of significant differences with respect to these variables is a statistical artifact. With respect to Spelling Ability the situation is less clear.

Results showed that the mean differences between the program outcomes ranged from .20 to .25. These differences are much smaller than the differences that were expected when the study was planned. According to the criteria of Cohen (1988), such differences should be considered small. The fact that we found only small and statistically nonsignificant differences is somewhat remarkable in the light of findings in the USA demonstrating learning differences between programs (Chall, 1967; Stahl & Miller, 1989). However, as we noted earlier, the didactical approach of beginning reading in The Netherlands is more or less monolithic. Every literacy program adheres to a phonics approach, more specifically the instruction of synthetic phonics. Students are explicitly taught to convert letters into sounds, and then blend the sounds to form recognizable words. It is our impression that the reading approach in the USA is much more variable, as shown by the continuing debate on phonics instruction (National Reading Panel, 2000, Chapter 2, part II). Our programs differed mainly with respect to their whole-class versus student orientation and the richness of teaching materials. It seems that these two variables are not related to the learning outcomes.

Do the findings of the present study settle the argument with respect to differential effects of our three literacy programs for initial reading? It can safely be concluded that if there are any *mean* differences, these differences are rather small. The present study is less conclusive with respect to the effects of literacy programs on the *variability* of individual outcomes. We already pointed to the increased variability of pupil abilities for the RtR program. Although not statistically reliable, it may still hold an important cue for future research. RtR is an individually tailored program. Each student is approached at his or her own individual level. Low achievers are taught by means of highly structured materials, whereas more able students are free to choose their own materials and work on their own. This is likely to result in increased differences between pupils within classes, providing a perfect instance of the so-called Matthew effect (Stanovich, 1986). To study these effects, substantially larger samples will be needed than the present one. It should be interesting to study the effects of literacy programs on interindividual variability of student outcomes.

A further suggestion would be to study the effects of selected contextual or classroom variables. Many contextual variables show only little variation, as The Netherlands is a small and relatively homogeneous country, and schools are equally treated by law, contributing to a nonsegregated educational system. National research shows many contextual variables to have little or no relationship with learning outcomes (Ledoux, Overmaat, & Koopman, 1997). One of the scarce exceptions is the experience of teachers. Teaching experience, expressed as the number of years a teacher teaches, shows a consistent positive relationship with learning outcomes. Specific attention to teacher characteristics, including his or her way of teaching, could provide further insight into the background of the school-level variability of learning outcomes.

REFERENCES

Baumann, J.F., & Heubach, K.M. (1996). Do basal readers deskill teachers? A national survey of educators' use and opinions of basals. *The Elementary School Journal*, *96*, 511–526.

Baumann, J.F., Hoffman, J.V., Moon, J., & Duffy-Hester, A.M. (1998). Where are teachers' voices in the phonics/whole language debate? Results from a survey of U.S. elementary classroom teachers. *The Reading Teacher*, *51*, 636–650.

Blok, H., Otter, M.E., & De Glopper, K. (2000). Vergelijkend onderzoek naar methoden voor aanvankelijk lezen [A comparative analysis of literacy programs for beginning reading instruction]. *Pedagogiek, Wetenschappelijk Forum voor Opvoeding, Onderwijs en Vorming*, *20*, 255–272.

Bond, G.L., & Dykstra, R. (1967). The cooperative research program in first-grade-reading instruction. *Reading Research Quarterly*, *2*, 1–142. (Reprinted in *Reading Research Quarterly*, *32*, 348–427.)

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Chall, J.S. (1967). *Learning to read: The great debate*. New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Duffy, G.G., & Hoffman, J.V. (1999). In pursuit of an illusion: The flawed search for a perfect method. *The Reading Teacher*, *53*, 10–16.

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*, 250–287.

Elley, W.B. (1992). *How in the world do students read*? The Hague, The Netherlands: The International Association for the Evaluation of Educational Achievement.

Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*, 37–55.

Foorman, B.R., Francis, D.J., Winikates, D., Mehta, P., Schatschneider, C., & Fletcher, J.M. (1997). Early interventions for children with reading disabilities. *Scientific Studies of Reading*, *1*, 255–276.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goodman, K.S., Shannon, P., Freeman, Y.S., & Murphy, S. (1988). *Report card on basal readers*. Katonah, NY: Richard C. Owen.

Kooreman, H.J. (1974). Konstruktie en resultaten van een onderwijsleerpakket voor het technisch leren lezen [The development and learning outcomes of a teaching program for initial reading instruction]. *Pedagogische Studiën*, *51*, 398–412.

Ledoux, G., Overmaat, M., & Koopman, P. (1997). Kwaliteitszorg in het primair onderwijs; secundaire analyses op de PRIMA-cohort bestanden [Quality management in primary education; secondary analyses of data from the PRIMA-cohort study]. Amsterdam: SCO-Kohnstamm Instituut.

Martinez, M.G., & McGee, L.M. (2000). Children's literature and reading instruction: Past, present, and future. *Reading Research Quarterly*, *35*, 154–169.

National Reading Panel. (2000). *Report of the National Reading Panel: Report of the subgroups*. Washington, DC: National Institute of Child Health and Human Development Clearinghouse. [http://www.nichd.nih.gov/publications/nrp/report.pdf]

Overmaat, M., & Ledoux, G. (1998). *School- en klaskenmerken basisonderwijs. Basisrapportage PRIMA-cohortonderzoek. Tweede meting 1996–1997* [Findings of the second PRIMA-cohort measurement round 1996–1997]. Amsterdam: SCO-Kohnstamm Instituut.

Searfoss, L.W. (1997). Connecting the past with the present: The legacy and spirit of the First Grade Studies. *Reading Research Quarterly*, *32*, 433–438.

Sijtstra, J. (1992). *Balans van het taalonderwijs halverwege de basisschool* [Results from the national survey regarding Dutch language education]. Arnhem, The Netherlands: Cito.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Stahl, S.A., & Miller, P.D. (1989). Whole language and language experience approaches for beginning reading: A quantitative research synthesis. *Review of Educational Research*, *59*, 87–116.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.

Stedman, L.C., & Kaestle, C.E. (1987). Literacy and reading performance in the United States from 1880 to the present. *Reading Research Quarterly*, *22*, 8–46.

Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs; The relationship between decoding and text. *Remedial and Special Education*, *20*, 275–287.

Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrica*, *58*, 395–415.

Williams, P.L., Reese, C.M., Campbell, J.R., Mazzeo, J., & Phillips, G.W. (1995). *1994 NAEP Reading: A first look – findings from the national assessment of educational progress*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.